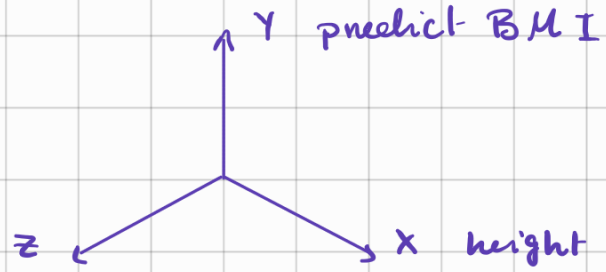


$y = \beta_0 + \beta_1 X + \beta_2 Z$

where X and Z are indep.

$y = \beta_0 + \beta_1 X^2 + \beta_2 Z^3$

multiple non linear regression

* Suppose $X_1, \ldots, X_k$ are independent :

$$y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + \hat{\varepsilon}_i \quad , \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

(random error)

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{2i} + \ldots + b_{k-1} x_{i\,k-1} + \hat{e}_i \quad , \quad n > k$$

(residual)

* Least Square Method :

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_{k-1} x_{i\,k-1})$$

1st normal equation

$$\frac{\partial L}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_{k-1} x_{i\,k-1}) = 0 = \sum e_i$$

$$\frac{\partial L}{\partial b_j} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_j x_{ij} - \ldots - b_{k-1} x_{i\,k-1}) x_{ij} = 0$$

j^{th} normal equation : $\sum e_i x_{ji} = 0$

$$\begin{cases} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_{k-1} x_{i\,k-1}) = 0 \\ \sum_{i=1}^{n} (y_i x_{ij} - b_0 x_{ij} - \ldots - b_j x_{ij}^2 - \ldots - b_{k-1} x_{i\,k-1} x_{ij}) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} b_0 = \bar{Y} - b_1 \bar{X}_1 - \cdots - b_k \bar{X}_k \\ b_j = \dfrac{\sum y_i x_{ij} - b_0 x_{ij} - \cdots - b_{j-1} x_{i\,\overline{j-1}} - b_{j+1} x_{i\,\overline{j+1}} - \cdots - b_{k-1} x_{i\,\overline{k-1}}\, x_{ij}}{\sum x_{ij}^2} \end{cases}$$

Thus, we get a system of K linear equations

* $\beta_0, \ldots, \beta_k$ are regression coefficients

* When estimating the parameters we only need to assume that $\mathbb{E}(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and they are iid but for testing we need to assume normality $\varepsilon_i \sim N(0, \sigma^2)$ otherwise F test fails.

* β_i represents the change in y_i when x_i changes and the other variables are constant so β_i's are called partial regression coefficients.

* **Least Square Method using Matrix Method :**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i\,\overline{k-1}}, \quad n > k-1$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{k-1} \end{pmatrix}_{k\times 1} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,k-1} \end{pmatrix}$$

$$Y = X\beta + \varepsilon$$

...riable ...atrix ... coeff ...ls ...r Form

$$\begin{pmatrix} \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK-1} \end{pmatrix}_{n \times K}$$

$E(\varepsilon) = 0$ vector $\qquad Var(\varepsilon) = \sigma^2 I$

$$E(Y) = E(X\beta + \varepsilon) = X\beta + E(\varepsilon) = X\beta$$
$$Var(Y) = Var(X\beta + \varepsilon) = Var(\varepsilon) = \sigma^2$$

Estimate of β by LSE :

$$\hat{Y} = X\hat{\beta} \qquad Y = X\hat{\beta} + \varepsilon$$

$$L = \sum_{i=1}^{n} e_i^2 = (e_1 \; \cdots \; e_n) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = e^T . e$$

$$= (Y - \hat{Y})^T (Y - \hat{Y})$$
$$= (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$
$$= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta})$$
$$= Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y - \hat{\beta}^T X^T X\hat{\beta}$$

as $c^T = c$ where c is a scalar $\leftarrow$
$$= Y^T Y - (\hat{\beta}^T X^T Y)^T - \hat{\beta}^T X^T Y - \hat{\beta}^T X^T X\hat{\beta}$$
$$= Y^T Y - 2\hat{\beta}^T X^T Y^T - \hat{\beta}^T X^T X\hat{\beta}$$

$$\frac{\partial}{\partial \hat{\beta}} = -2X^T Y - 2X^T X\hat{\beta} = 0 \Rightarrow X^T X\hat{\beta} = X^T Y$$
$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

Hence, $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$

When $K = 1$, we get back to simple linear regression.

* In the chocolate company example, some other regressors

could be the number of sales person, etc. In multiple linear regression we have $k-1$ regressors; $x_{i1}, \dots, x_{i\,k-1}$

i	x_{i1}	x_{i2} -----	$x_{i\,k-1}$	$\det(X^T X) = 0$ not invertible and the colums are linearly dependent otherwise they are independent and $\det(X^T X) \neq 0$ so some features might be redundant
1	x_{11}	x_{12} ----	$x_{1\,k-1}$	
2	x_{21}	x_{22} -----	$x_{2\,k-1}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	x_{n1}	x_{n2} -----	$x_{n\,k-1}$	

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}\left((X^T X)^{-1} X^T Y\right) \\
&= (X^T X)^{-1} X^T \mathbb{E}(X\beta + \varepsilon) \\
&= (X^T X)^{-1} X^T \left(X\beta + \mathbb{E}(\varepsilon)\right), \quad \mathbb{E}(\varepsilon) = 0 \\
&= (X^T X)^{-1} (X^T X)\beta \\
&= \beta \quad \text{unbiased estimator}
\end{aligned}
$$

$$
X^T = \begin{pmatrix}
1 & 1 & \dots & 1 \\
x_{11} & x_{21} & \dots & x_{n1} \\
x_{12} & x_{22} & \dots & x_{n2} \\
\vdots & \vdots & & \vdots \\
x_{1\,k-1} & x_{2\,k-1} & \dots & x_{n\,k-1}
\end{pmatrix}_{k \times n}
$$

$X^T X$ is symmetric, we call it the variance - covariance matrix.

* **Gauss Markov Theorem**:

The OLS (ordinary least square) estimates $\hat{\beta}$ of β is the best linear unbiased estimator (BLUE).

* **Proof**: Elements of Statistical Learning
Linear Regression Analysis by Montgomery

* **Def**: BLUE

An estimateur $\hat{\theta}$ is called BLUE for θ if

1) $\hat{\theta}$ is a linear combination of samples obs ($\hat{\beta}_i = \sum c_i y_i$ is linear)
2) $Var(\hat{\theta}) \leq Var(\theta')$ where $\hat{\theta}$ and θ' are two unbiased estimateurs.

$$
\begin{aligned}
Var(\hat{\beta}) &= Var\left((X^T X)^{-1} X^T Y\right) \\
&= \left((X^T X)^{-1} X^T\right)\left((X^T X)^{-1} X^T\right)^T Var(Y) \\
&= (X^T X)^{-1} X^T X (X^T X)^{-1} \sigma^2 I \\
&= I \cdot (X^T X)^{-1} \sigma^2 I \\
&= \sigma^2 (X^T X)^{-1} \\
&= MS_{Res} (X^T X)^{-1} \qquad \text{when } \sigma \text{ unknown, } n-K \text{ deg.}
\end{aligned}
$$

$$SS_{Res} = \sum e_i^2 \sim N(0, \sigma^2) \quad \Rightarrow \quad \frac{e_i}{\sigma} \sim N(0,1)$$

$$\Rightarrow \quad \frac{e_i^2}{\sigma^2} \sim \chi_1^2$$

because we have K normal eq. $$\Rightarrow \quad \frac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-K}$$

$E(MS_{Res}) = \sigma^2$

$$SS_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$ $n-1$ df as we have one restriction $\sum Y_i - \bar{Y} = 0$

$$MS_{Res} = \frac{SS_{Res}}{n-K}$$ Then, SSreg has $K-1$ degrees of freedom

* **Example**: In R.

1) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, compute β_i's $i = 0, 1, 2$

2) Find Anova table

Source of variability	DF	SS	MS	F-value
Regression	$K-1$	$\sum (\hat{Y} - \bar{Y})^2$	$SSreg / (k-1)$	$F = \frac{MSreg}{MSres} \sim F_{K-1, n-K}$
Residual	$n-k$	$\sum (Y_i - \hat{Y})^2$	$SSres / (n-k)$	
Total	$n-1$	$\sum (Y_i - \bar{Y})^2$	$SS_T / (n-1)$	

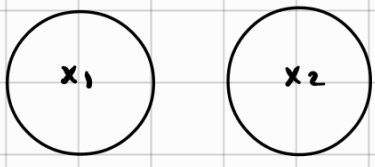
$F_{obs} > F_{1, n-K, 0.05}$

$H_0 : \beta_1 = \beta_2 = 0$ rejected otherwise MLR is not useful

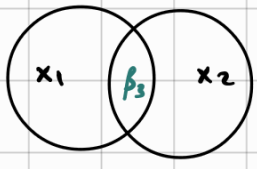
3) Calculate R^2 $\quad R^2 = \frac{SS_{reg}}{SS_T}$

4) Calculate the estimated variance of $\hat{\beta}$.

* Interactions :



$Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \varepsilon$



$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

* Global Test :

$H_0 : \beta_j = 0 \quad j = 1(1)K-1$ vs $H_1 : \beta_j \neq 0 \quad \exists j \in \{0, \dots, K-1\}$

$\frac{SSRes}{\sigma^2} \sim \chi^2_{n-K}$ $\qquad$ $\frac{MSReg}{MSRes} \sim F_{K-1, n-K}$

$\frac{SSReg}{\sigma^2} \sim \chi^2_{K-1}$ $\qquad$ reject H_0 if $F_{cal} > F_{\alpha, K-1, n-K}$ i.e $\exists \beta_j \neq 0$

* Partial Test : Which regressor is significant ?

* Partial Test : which regressor is significant ?

$H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ $\quad j = 1(1)K-1$

$\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj}) \Rightarrow \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}} \sim N(0, 1)$

under H_0 : $\dfrac{\hat{\beta}_j}{\sqrt{MSRes\,(X^T X)^{-1}_{jj}}} \sim t_{\alpha/2, n-K}$

If $|t| > t_{\alpha/2, n-K}$, we reject H_0 in this case we do not need to do a global test as $\exists \beta_j \neq 0$

* Confidence Interval : $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{MSRes\,(X^T X)^{-1}_{jj}}} \sim t_{\alpha/2, n-K}$

$$\mathbb{P}\left(\left| \frac{\hat{\beta}_j - \beta_j}{\sqrt{MSRes\,(X^T X)^{-1}_{jj}}} \right| < t_{\alpha/2, n-K} \right) = 1 - \alpha$$

$$\mathbb{P}\left(\beta_j - t_{\alpha/2, n-K} \sqrt{MSRes\,(X^T X)^{-1}_{jj}} < \hat{\beta}_j < \beta_j + t_{\alpha/2, n-K} \sqrt{MSRes\,(X^T X)^{-1}_{jj}} \right) = 1 - \alpha$$

* Proof of Gauss Markov Theorem :

We want to show the OLS is the best linear unbiased estimate of β.

By OLS : $\hat{\beta} = (X^T X)^{-1} X^T Y$

$\mathbb{E}(\hat{\beta}) = \beta \qquad Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

Now, consider another unbiased estimator of β, call it $\tilde{\beta}$. We use the fact that any other estimator of β can be written as

$$\tilde{\beta} = \left((X^TX)^{-1}X^T + B\right)Y + B_0, \quad B \in \mathcal{M}_{k\times n}(\mathbb{R}), \quad B_0 \in \mathcal{M}_{n\times 1}(\mathbb{R})$$

$$\begin{aligned}
\mathbb{E}(\tilde{\beta}) &= \mathbb{E}\left((X^TX)^{-1}X^TY\right) + \mathbb{E}(\beta Y) + \mathbb{E}(\beta_0) \\
&= \beta + B\,\mathbb{E}(X\beta + \epsilon) + B_0 \\
&= \beta + BX\beta + B_0 \qquad , \beta \in \mathcal{M}_{k\times n}(\mathbb{R})
\end{aligned}$$

Note that BX and B_0 are zeros as we assumed that $\tilde{\beta}$ is unbiased

$$\begin{aligned}
Var(\tilde{\beta}) &= Var\left(\left((X^TX^{-1})X^T + B\right)Y + \beta_0\right) \\
&= Var\left(\left((X^TX^{-1})X^T + B\right)Y\right) \\
&= \left((X^TX^{-1})X^T + B\right)\left((X^TX^{-1})X^T + B\right)^T Var(Y) \\
&= \left((X^TX)^{-1}X^T + B\right)\left(X(X^TX)^{-1} + B^T\right)\sigma^2 \\
&= \sigma^2\left((X^TX)^{-1} + (X^TX)^{-1}(BX)^T + (BX)(X^TX)^{-1} + \beta\beta^T\right) \\
&= \sigma^2\left((X^TX)^{-1} + \beta\beta^T\right)
\end{aligned}$$

Therefore, $Var(\tilde{\beta}) > Var(\hat{\beta})$ and thus OLS estimators of β is the BLUE because $\beta\beta^T > 0$